April 1965

U.S. FOREST SERVICE
RESEARCH NOTE RM - 44

FOREST SERVICE
U.S. DEPARTMENT OF AGRICULTURE

ROCKY MOUNTAIN FOREST AND RANGE EXPERIMENT STATION

# Multiple Comparison Procedures

Martin A. Hamilton [1]

The standard method of comparing $k$ experimental treatment means, $\overline{X}_i$, $i = 1, \ldots, k$, is to construct an analysis of variance table and conduct an $F$ test. If the $F$ test is significant, the experimenter can only state that all the means are not equal. But he usually wants to ask more specific questions about differences among treatment means than can be answered by this $F$ test. In fact, the experimenter often wishes to decide which of the true treatments means, $\mu_i = E(\overline{X}_i)$, differ from each other.

Many methods are proposed in statistical literature for comparing experimental treatment means (Federer 1955; Hartley 1955; Scheffé 1959).[2] This paper is a review of that literature and an attempt at a unified presentation of multiple comparison methods for workers in the field.

The plan of this paper is to first provide a foundation on which different multiple comparison methods may be presented, and then to describe and compare some well-known procedures. The first section, therefore, discusses the types of α-error rates (levels of significance) provided by various multiple comparison tests, and defines terms that are used in subsequent sections. A general multiple comparison procedure is given in section 2. In section 3, five multiple comparison tests are individually described, and hints are given for selecting the appropriate test for a particular problem. Section 4 contains an example problem illustrating all five methods. The fifth and final section defines two methods of testing multiple contrasts among treatment means, discusses their applicability, and analyzes example data using both procedures. For most forestry problems, the field worker will find that sections 1.3, 1.5, 2, 3.3, 3.4, 3.6, 4.3, and 4.4 are of the greatest value.

## 1. α-ERROR RATES AND POWER

The experimenter must attach some nominal level of significance to his statements concerning differences between treatment means. If he has only two means to compare, the usual "$t$-test" at

[1] *Mathematical Technician during summer of 1964 at Rocky Mountain Forest and Range Experiment Station central headquarters at Fort Collins, in cooperation with Colorado State University; now a graduate student and candidate for Ph. D. degree, Statistics Department, Stanford University, Stanford, California.*

[2] *Names and dates in parentheses refer to Literature Cited, page 12 .*

5

the $\alpha$ level of significance will yield a $100\alpha$ percent chance of stating that the true means are different when they are actually equal. But if the experimenter has $k$ experimental means to compare pairwise, he is simultaneously testing ${}_kC_2 = k(k-1)/2$ null hypotheses of the form $H_0: \mu_i = \mu_j$ against the alternate hypotheses $H_1: \mu_i \neq \mu_j$ (e.g., if $k = 4$, the experimenter may want to make six inferences at the same time about the values of $\mu_4 - \mu_1$, $\mu_4 - \mu_2$, $\mu_4 - \mu_3$, $\mu_3 - \mu_1$, $\mu_3 - \mu_2$, and $\mu_2 - \mu_1$). An experiment yielding $k$ means thus may require as many as $k(k-1)/2$ simultaneous "inferences." The $\alpha$ error (level of significance) for this case has been given more than one meaning (Hartley 1955). The following definitions of error will be encountered in this paper.[3]

### 1.1 Error Rate Per Comparison

$\alpha_c$ = error rate per comparison = (no. of erroneous inferences) / (no. of inferences attempted) = proportion of all comparisons expected to be erroneous when the null hypotheses are true. For example, with 100 experiments, each yielding 20 treatment means and 190 paired comparisons, the experimenter expects to wrongly reject 950 of the 19,000 null hypotheses if he uses an $\alpha_c$ error rate of .05.

This error rate is appropriate when:

a. The specific comparison is the conceptual unit.
b. The proportion of all erroneous inferences to all comparisons made is to be a constant.
c. The experimental error variance is relatively stable from experiment to experiment (see section 1.5).
d. A faulty inference does not affect the remaining inferences from the experiment (see section 1.5).

### 1.2 Error Rate Per Experiment

$\alpha_e$ = error rate per experiment = (no. of erroneous inferences) / (no. of experiments) = the expected number of erroneous inferences per experiment when the null hypotheses are true. For example, with 100 experiments, each yielding 20 treatment means and 190 paired comparisons, the experimenter expects to wrongly reject 9 or 10 of the 190 null hypotheses in <u>each</u> of the 100 experiments if he uses an $\alpha_e$ error rate of .05.

This error rate is appropriate when:

a. The experiment is the conceptual unit.
b. The average number of erroneous inferences per experiment is to be kept constant.
c. The experimental error variance is relatively stable from experiment to experiment (see section 1.5).
d. A faulty inference for one comparison does not affect the remaining comparisons (see section 1.5).

### 1.3 Experimentwise[4] Error Rate

$\alpha_w$ = experimentwise error rate = (no. of experiments with one or more erroneous inferences) / (no. of experiments) = expected proportion of experiments with one or more erroneous inferences when the null hypotheses are true. For example, with 100 experiments, each yielding 20 treatment means and 190 paired comparisons, the experimenter expects to wrongly reject one or more null hypotheses in <u>only</u> 5 of the 100 experiments if he uses an $\alpha_w$ error rate of .05.

This error rate is appropriate when:

a. The experiment is the conceptual unit.
b. The average proportion of experiments in which <u>one or more</u> faulty inferences are made is to be kept constant.
c. The experimental error variance fluctuates from experiment to experiment (see section 1.5).
d. The value of other inferences from the experiment is lowered as soon as <u>one</u> faulty inference is made (see section 1.5).

[3] *Federer, W. T. Error rates in experiments. (Lecture notes from Advanced Science Seminar in Mathematical Statistics, Dept. of Math. and Statis., Colo. State Univ., Ft. Collins, Colo., Aug. 7, 1964.)*

[4] *"Experimentwise" is the term suggested by J. W. Tukey in the early 1950's to identify this type of error rate.*

## 1.4 Duncan's Protection Level

D. B. Duncan (1955) defined the protection level concept for use in a particular type of multiple comparison test (see sections 3.5 and 4.5). This concept is difficult to explain in terms parallel to those used in sections 1.1, 1.2, and 1.3. For this reason, discussion of the protection level is limited to an interpretation of its use in Duncan's New Multiple Range Test.

An experiment is said to be of "type $d$," $d = 1, 2, \ldots, k$ if:

i) It yields $k$ experimental means, where $k \geq d$.

ii) $d$ of the $k$ true treatment means are equal to $\mu$.

iii) The remaining $k$-$d$ true treatment means are all different from $\mu$ and all different from each other.

In other words, an experiment is of "type $d$" if there is only one cluster of associated true treatment means, and this cluster is of size $d$.

Now let $\alpha_p$ = the level of significance the experimenter would choose for a test of the difference between any two means, assuming that the remaining means were not present. Then Duncan's protection level for an experiment of "type $d$" $= 1 - (1 - \alpha_p)^{d-1}$ = (no. of experiments of type $d$ with one or more erroneous inferences) / (no. of experiments of type $d$). For example, with 100 experiments, each yielding 20 treatment means and 190 paired comparisons, the experimenter expects to wrongly reject one or more null hypotheses in 62 of the experiments if he uses an $\alpha_p$ error rate of .05 and if each of the 100 experiments is of type 20. If each of the experiments is of type 2, the experimenter expects to wrongly reject one or more null hypotheses in 5 of the experiments when he uses an $\alpha_p$ error rate of .05.

This error rate is appropriate when:

a. The experiment is the conceptual unit.

b. The average proportion of experiments of type $d$ in which one or more faulty inferences are made is to be kept constant.

c. The experimental error variance fluctuates from experiment to experiment (see section 1.5).

d. The value of other inferences from an experiment of type $d$ is lowered as soon as one faulty inference is made, but the amount that the value is lowered is small if $d$ is large (see section 1.5).

## 1.5 Discussion of $\alpha$ Error Rates

By comparing conditions $c$ and $d$ under sections 1.1, 1.2, 1.3, and 1.4, it seems reasonable to conclude that in most forestry experiments an experimentwise error rate, $\alpha_w$, should be used (Hartley 1955). Errors of experimentation usually do affect the entire experiment. This is certainly true when the experimenter uses a multiple comparison procedure to help him choose some preferred treatments. If even one false inference is made, the experiment provides little information about the true relationships among treatment means. In forestry research, experimental error variance fluctuations from experiment to experiment are common. For example, varying climatic conditions affect the variance, but are impossible to control from experiment to experiment. For these reasons, this paper is slanted toward the use of experimentwise error rates. The experimenter must remember to consider all four definitions of error rate, however, and choose the one that is most appropriate for his situation.

## 1.6 The Terms "Conservative" and "Powerful"

The power of a test is defined as the probability of rejecting the null hypothesis when it is false. Power is a function of $\mu_i - \mu_j$, the actual mean differences, of $\sigma^2$, the common variance, and of $\alpha$, the level of significance. If two statistical tests are available for analyzing a set of data, the experimenter should use the test that guarantees the higher power for the error rate he has chosen.

Suppose the experimenter wishes to compare a test based on an error rate per comparison with a test based on an experimentwise error rate. It can be shown that if $\alpha_c = \alpha_w$, the test based on an error rate per comparison is more powerful. But this is not a true power comparison because the error rates, $\alpha_c$ and $\alpha_w$, are not the same by definition. Although statisticians have compared them, the power functions of tests based on differently defined error rates should not be directly compared. For this reason, the term "conservative" is used in the following discussion. The statement "Fisher's test is more conservative than Tukey's test" indicates that Tukey's test is more powerful in the sense that the powers are compared at $\alpha_e = \alpha_w$.

A simple way to decide which of two tests is more conservative (more powerful if the tests are based on the same error rate) is to compare the lengths of the $1 - \alpha_c$, $1 - \alpha_e$, $1 - \alpha_w$, and $1 - \alpha_p$ confidence intervals derived from the tests, where $\alpha_c = \alpha_e = \alpha_w = \alpha_p$. The test producing the wider confidence intervals about differences between pairs of means is the more conservative.

## 2. GENERAL MULTIPLE COMPARISON TEST PROCEDURE

Let all $k$ experimental means be based on $n$ observations and have common unknown variance $\sigma^2/n$. The general multiple comparison procedure is to order the experimental means from smallest to largest, assigning consecutive indices so that $\overline{X}_1 < \overline{X}_2 < \ldots < \overline{X}_{k-1} < \overline{X}_k$. Then a table (see table 1) of differences of all possible pairs is formed. Each table entry is compared with an appropriate critical value $K(\alpha, v)\ s/\sqrt{n}$, where $s^2$ is an independent estimate of $\sigma^2$ based on $v$ degrees of freedom and $K(\alpha, v)$ is a tabular value. If $\overline{X}_i - \overline{X}_j$ is greater than $K(\alpha, v)\ s/\sqrt{n}$, the null hypothesis, $H_0$: $\mu_i = \mu_j$, is rejected and the alternate hypothesis, $H_1$: $\mu_i \neq \mu_j$, is accepted. The difference, $\overline{X}_i - \overline{X}_j$, is then said to be "significant." This is the same testing procedure as for the ordinary $t$-test, but with different tabled values.

Table 1--Differences between pairs of means

| | $\overline{X}_1$ | $\overline{X}_2$ | ... | $\overline{X}_{k-1}$ | $\overline{X}_k$ |
|---|---|---|---|---|---|
| $\overline{X}_k$ | $\overline{X}_k - \overline{X}_1$ | $\overline{X}_k - \overline{X}_2$ | ... | $\overline{X}_k - \overline{X}_{k-1}$ | 0 |
| $\overline{X}_{k-1}$ | $\overline{X}_{k-1} - \overline{X}_1$ | $\overline{X}_{k-1} - \overline{X}_2$ | ... | 0 | |
| $\vdots$ | $\vdots$ | $\vdots$ | | | |
| $\overline{X}_2$ | $\overline{X}_2 - \overline{X}_1$ | 0 | | | |
| $\overline{X}_1$ | 0 | | | | |

Multiple comparisons based on an $\alpha_w$ error rate are sometimes better presented as simultaneous confidence intervals, where the probability is greater than or equal to $1-\alpha$ that simultaneously for all differences between pairs of treatment means,

$$\overline{X}_i - \overline{X}_j - K(\alpha,\ v)\ s/\sqrt{n} \leq \mu_i - \mu_j \leq \overline{X}_i - \overline{X}_j + K(\alpha,\ v)\ s/\sqrt{n}.$$

Simultaneous confidence intervals may, in turn, be regarded as tests of hypothesis if the null hypothesis $H_o$: $\mu_i = \mu_j$ is rejected when and only when the interval $\overline{X}_i - \overline{X}_j \pm K(\alpha, v)\ s/\sqrt{n}$ does not contain zero.

The experimenter should try to define his problem and choose an appropriate method of analysis before he looks at the data. If tests of differences between all pairs of experimental means are paramount, the multiple comparison procedure is to be applied directly--not after an $F$-test. Some experimenters use a multiple comparison procedure only if the $F$-test is significant. This is not an efficient approach. The $F$-test does not help solve the problem, but only succeeds in making the multiple comparison test more conservative.

The standard procedure is to calculate the experimental means, form an analysis of variance table, use the mean square error as an independent estimate of the variance, and apply the appropriate multiple comparison test.

## 3. SOME MULTIPLE COMPARISON TESTS

There are two types of multiple comparison tests. With one type, all differences, $\bar{X}_i - \bar{X}_j$, are compared with the same critical value. With the other type the differences, $\bar{X}_i - \bar{X}_j$, are compared with critical values that depend upon the number $i-j$. These procedures are called fixed range tests and multiple range tests, respectively.

The five tests discussed in this paper do not exhaust the literature on multiple comparison methods. These tests are, however, the best known and/or the most useful procedures.

### 3.1 L.S.D. or Multiple $t$-test; $\alpha_c$ Error Rate; Fixed Range

This test is conducted by letting $K(\alpha, v) = \sqrt{2}\ t(\alpha, v)$, where $t(\alpha, v)$ is the upper $100\alpha/2$ percent point of the $t$ distribution based on $v$ degrees of freedom (Federer 1955). The differences $\bar{X}_i - \bar{X}_j$, are compared with the least significant difference (L.S.D.), $\sqrt{2}\ t(\alpha, v)\ s/\sqrt{n} = K(\alpha, v)\ s/\sqrt{n}$, as described in the general test procedure.

The $\alpha$ error associated with the L.S.D. test is error rate per comparison $\alpha_c$. If the experimentwise error rate is important, this test can be misleading. The following tabulation shows that if 20 means are to be compared, the L.S.D. test based on $t(.05, v)$ actually allows an $\alpha_w$ error of 90 percent:

| $k$ = number of means | experimentwise error |
|---|---|
| 2 | .05 |
| 6 | .34 |
| 12 | .68 |
| 15 | .72 |
| 20 | .90 |

Needless to say, this is not a satisfactory procedure to use if we wish to control on experimentwise error rate.

The L.S.D. test was the first method suggested to solve the multiple comparison problem, but recently has found limited acceptance among experimenters.

### 3.2 Fisher's Test; $\alpha_e$ Error Rate; Fixed Range

Fisher's test is conducted exactly as the L.S.D. test, except that it is adjusted to insure an error rate per experiment, $\alpha_e$ (Harter 1957). This is accomplished by letting $K(\alpha, v) = \sqrt{2}\cdot t(\alpha/m, v)$ where $m$ is the number of paired comparisons desired by the experimenter. Usually, $m = k(k-1)/2$, the total number of possible pairs of $k$ experimental means. The general procedure is followed using the critical value $\sqrt{2}\ t(\alpha/m, v)\ s/\sqrt{n} = K(\alpha, v)\ s/\sqrt{n}$. Fisher's test is the most conservative procedure mentioned in this paper.[5] In fact, if $k = 6$, $n = 25$, and $\alpha_c = .067$, the corresponding experimentwise error rate is $\alpha_w = .05$ (Harter 1957).

### 3.3 Tukey's Test; $\alpha_w$ Error Rate; Fixed Range

Many textbooks (Federer 1955, pp. 22-23; Scheffé 1959, pp. 434-436) contain tables of the upper percentage points, $q(\alpha, k, v)$, of the Studentized range for various values of $k$, $v$, and $\alpha$. Tukey's test consists of finding $q(\alpha, k, v)$ in tables of the Studentized range, and comparing the differences $\bar{X}_i - \bar{X}_j$ with $q(\alpha, k, v)\ s/\sqrt{n} = K(\alpha, v)\ s/\sqrt{n}$ as in the general procedure. The experimentwise error, $\alpha_w$, for this method is guaranteed to be less than or equal to the $\alpha$ in $q(\alpha, k, v)$. There are other fixed-range tests that control on experimentwise error rates (Scheffé 1959), but Tukey's test is the most powerful.

### 3.4 Newman-Keuls' Test; $\alpha_w$ Error Rate; Multiple Range

This test also depends upon the Studentized range but is less conservative than Tukey's test (Hartley 1955). The standard procedure is as follows:

[5] *Dunn (1961) has shown exceptions to this statement in certain well-defined situations.*

a. $\overline{X}_i - \overline{X}_j$ is said to significant if
   $\overline{X}_i - \overline{X}_j > q(\alpha, i-j+1, v)\ s/\sqrt{n}$, where $q(\alpha, i-j+1, v)$ is the upper 100$\alpha$ percent point of the studentized range for $i-j+1$ observed means and $v$ degrees of freedom.
b. Construct the table of mean differences as illustrated in section 2.
c. Begin testing in the upper lefthand cell. If $\overline{X}_k - \overline{X}_1$ is not significant, state that all the means are equal and conclude the test. If $\overline{X}_k - \overline{X}_1$ is significant, test $\overline{X}_k - \overline{X}_2$ and $\overline{X}_{k-1} - \overline{X}_1$.
d. If $\overline{X}_k - \overline{X}_2$ is not significant, state that $\overline{X}_2 = \overline{X}_3 = \ldots = \overline{X}_k$ and test no further differences in the 2nd through $k$th columns. If $\overline{X}_k - \overline{X}_2$ is significant, test $\overline{X}_k - \overline{X}_3$.
e. Continue in this manner, testing a difference, $\overline{X}_i - \overline{X}_j$, only if significant differences have been found in all cells above and left of the cell containing $\overline{X}_i - \overline{X}_j$.

If there is only one "cluster" of true means, the experimentwise error is guaranteed to be less than or equal to $\alpha$. But if there are $m$ clusters of true means, $\alpha_w$ is only guaranteed to be less than or equal to $m\alpha$. A cluster is defined as two or more equal true means. For example, if the true relation among the means is $\mu_1 = \mu_2 < \mu_3 = \mu_4 < \ldots < \mu_{k-1} = \mu_k$ so that there are $k/2$ clusters of true means, the experimentwise error rate could be as high as $(k/2)\alpha$, which is certainly a limitation of this procedure.

### 3.5 Duncan's New Multiple Range Test; $\alpha_p$ Error Rate; Multiple Range

This test procedure is conducted exactly as the Newman-Keuls' test except that the tabled values, $q(\alpha, i-j+1, v)$, have been revised to provide a less conservative test (Duncan 1955). To use this method, the revised values, $q^*(\alpha, i-j+1, v)$, must be read from special tables, usually titled "Critical Values for Duncan's New Multiple Range Test" (Federer 1955; Harter 1960b). These values are slightly smaller than corresponding values in a standard Studentized range table. Duncan justifies this method by defining the protection level concept and argues that the experimentwise error rate is too strong a criterion. To illustrate the difference, if $k = 6$ and $n = 25$, Duncan's test must be run with tabled values of $q^*(.01, i-j+1, v)$ to insure an experimentwise error of .05 (Harter 1957).

Duncan's test is affected in a more complicated manner than the Newman-Keuls' test if there are $m > 1$ clusters of the true means. This fact does not seem to be mentioned in the literature.

### 3.6 Selecting an Appropriate Test

In the preceding discussion, the term conservative was freely used. Power is a more meaningful concept to consider when selecting the better of two tests. To find a true comparison of the powers of these five multiple comparison tests, each method can be adjusted to yield the same experimentwise error rate. If this is done, computations indicate that all five tests have roughly the same power.

Since the powers are much the same, the choice of a test is actually governed by the choice of an appropriate $\alpha$ error rate. By considering the definitions and discussion of error rates given in section 1, the experimenter can avoid the confusion previously associated with the selection of a multiple comparison test. Note that if $k = 2$, all five methods are identical in both error rate and critical value.

Because the experimental error variance fluctuates from experiment to experiment in forestry work, the $\alpha_c$ and $\alpha_e$ error rates are not adequate in most cases. This fact generally eliminates the L.S.D. test and Fisher's test from consideration.

If it is assumed that the experimenter believes the true treatment means are not grouped in more than one cluster, how does he choose between Tukey's test, Newman-Keuls' test, and Duncan's test? An answer to this question depends upon the individual philosophy of the experimenter.

If he feels that the importance of a false inference diminishes as the number of true treatment means in the cluster increases, he will use the Duncan procedure. In using Duncan's test, he realizes that the probability of falsely rejecting one or more hypotheses is somewhere between $\alpha_p$ and $1-(1-\alpha_p)^{k-1}$. For the example in section 1.4 where $k = 20$, $\alpha_p = .05$, and $1-(1-\alpha_p)^{k-1} = .62$.

If the experimenter decides that one false inference decreases the value of the other inferences, regardless of the number of means in the cluster, he wants to control on an $\alpha_w$ error rate. This is usually the case in forestry research. Now either Newman-Keuls' test or Tukey's test can be chosen. The experimenter should realize that if he believes the Newman-Keuls' test at $\alpha_w = .05$ is too conservative, he can use a larger $\alpha_w$ error rather than adopting Duncan's procedure.

It is true that the experimenter can not know the relationship among the true treatment means. His guess as to the number of clusters of true treatment means is based on his experience and knowledge of the experimental situation. The statisticians can only state that if experimentwise error rate is important and the experimenter feels that the true means are not grouped in more than one cluster, the Newman-Keuls' test is appropriate. If the experimenter suspects that the true means are grouped in two or more clusters, he should use Tukey's test.

An experimenter can provide appropriate answers a large percent of the time if he routinely applies Tukey's test to every multiple comparison problem he encounters. This is true because generally experimentwise error is important, and because Tukey's test is valid no matter how the true means are grouped.

Sometimes the experimenter wishes to compare all treatments with a standard or a control. He may feel that only differences between treatment means and the control mean are of interest. The Newman-Keuls' test is appropriate for this situation. Dunnett (1955) proposes a fixed-range test for this case that requires a special set of tables.

### 3.7 Means Based on Unequal Numbers of Observations

One of the assumptions made in section 2 was that the experimental means are all based on the same number of observations. If this is not true, (i.e. $\bar{X}_i$ is based on $n_i$ observations and has a variance of $\sigma^2/n_i$) the tests must be adjusted (Sarhan and Greenberg 1962).

To adjust the L.S.D. test and the Fisher's test, multiply the usual tabled value by $s(1/n_i + 1/n_j)^{1/2}$ to test the difference $\bar{X}_i - \bar{X}_j$. The errors, $\alpha_c$ and $\alpha_e$, are not affected.

To adjust Tukey's test, Newman-Keuls' test, and Duncan's test, multiply the usual tabled values by $\sqrt{s^2/2(1/n_i + 1/n_j)}$ to test the difference $\bar{X}_i - \bar{X}_j$. This is a conservative approximation for all three methods (Kramer 1956).

## 4. EXAMPLE

Hartley (1955) quotes data on the comparative yields of naphthalene dye stuff prepared from $k = 6$ different "treatments" of H-acid. Each treatment was replicated $n = 5$ times. The means, arranged in ascending order, are given below.

| Treatment index, $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Treatment mean, $\bar{X}_i$ | 1,470 | 1,498 | 1,505 | 1,528 | 1,564 | 1,600 |

The analysis of variance shown below provides an estimate of the variance, $s^2 = 2,451$, with $v = 24$ degrees of freedom:

| | Degrees of freedom | Sum of squares | Mean square |
|---|---|---|---|
| Due to: | | | |
| Between treatments | 5 | 56,360 | 11,272 |
| Within treatments | 24 | 58,824 | 2,451 |
| Total | 29 | 115,184 | |

Next a table of mean differences is formed (table 2).

Table 2--Differences between experimental means

| | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ |
|---|---|---|---|---|---|---|
| $\bar{X}_6$ | 130 | 102 | 95 | 72 | 26 | 0 |
| $\bar{X}_5$ | 94 | 66 | 59 | 36 | 0 | |
| $\bar{X}_4$ | 58 | 30 | 23 | 0 | | |
| $\bar{X}_3$ | 35 | 7 | 0 | | | |
| $\bar{X}_2$ | 28 | 0 | | | | |
| $\bar{X}_1$ | 0 | | | | | |

### 4.1 L.S.D. Method - $\alpha_c = .05$

$t(.05, 24) = 2.06$;  $\sqrt{2}\ t(\alpha, v)\ s/\sqrt{n} = \sqrt{2}(2.06)\ (22.14) = 64.6$

The L.S.D., 64.6, is compared with the values in table 2 and the results are conveniently summarized by

$$\begin{array}{cccccc} \bar{X}_1 & \bar{X}_2 & \bar{X}_3 & \bar{X}_4 & \bar{X}_5 & \bar{X}_6 . \\ \cline{1-4} \cline{5-6} & & & & & \\ \cline{3-5} \end{array}$$

Any two means not underscored by the same line are significantly different.
Any two means underscored by the same line are not significantly different, as defined in section 2.

### 4.2 Fisher's Test - $\alpha_e = .05$

$t(.05/15, 24) = t(.0033, 24) = 3.429$,   $\sqrt{2}\ t(\alpha, v)\ s/\sqrt{n} = \sqrt{2}(3.429)\ (22.14) = 107.0$

The critical value, 107.0, is compared with the mean differences in table 2, and the results are summarized as in 4.1.

$$\begin{array}{cccccc} \bar{X}_1 & \bar{X}_2 & \bar{X}_3 & \bar{X}_4 & \bar{X}_5 & \bar{X}_6 \\ \cline{2-6} & & & & & \\ \cline{1-5} \end{array}$$

### 4.3 Tukey's Test - $\alpha_w = .05$

From tables of the Studentized range, $q(\alpha, k, v) = q(.05, 6, 24) = 4.37$.
$K(\alpha, v)\ s/\sqrt{n} = (4.37)\ (22.14) = 96.8$

The critical value, 96.8, is compared with the mean differences in table 2, and the results are summarized as in 4.1.

$$\begin{array}{cccccc} \bar{X}_1 & \bar{X}_2 & \bar{X}_3 & \bar{X}_4 & \bar{X}_5 & \bar{X}_6 \\ \cline{3-6} & & & & & \\ \cline{1-5} \end{array}$$

### 4.4 Newman-Keuls' Test - $\alpha_w = .05$

To find the critical value for the $\bar{X}_6 - \bar{X}_1$ cell, find $q(.05, 6-1+1, 24) = 4.37$ in the Studentized range table and multiply by $s/\sqrt{n} = 22.14$, so that $4.37(22.14) = 96.8$.

To find the critical value for the $\bar{X}_5 - \bar{X}_2$ cell, find $q(.05, 5-2+1, 24) = 3.90$ in tables of the Studentized range and multiply by $s/\sqrt{n}$ , so that $3.90(22.14) = 86.3$.

The test was concluded when $\bar{X}_6 - \bar{X}_4$, $\bar{X}_5 - \bar{X}_2$, and $\bar{X}_4 - \bar{X}_1$ were all found nonsignificant. The results are summarized as in 4.1.

$$\begin{array}{cccccc} \bar{X}_1 & \bar{X}_2 & \bar{X}_3 & \bar{X}_4 & \bar{X}_5 & \bar{X}_6 \\ \cline{4-6} & & & & & \\ \cline{2-5} & & & & & \\ \cline{1-4} \end{array}$$

Table 3--Critical Values for Newman-Keuls' Method with $\alpha = .05$, $K = 6$, $n = 24$

| | $\overline{X}_1$ | $\overline{X}_2$ | $\overline{X}_3$ | $\overline{X}_4$ | $\overline{X}_5$ | $\overline{X}_6$ |
|---|---|---|---|---|---|---|
| $\overline{X}_6$ | 96.8* | 92.3* | 86.3* | 78.2 | | |
| $\overline{X}_5$ | 92.3* | 86.3 | | | | |
| $\overline{X}_4$ | 86.3 | | | | | |
| $\overline{X}_3$ | | | | | | |
| $\overline{X}_2$ | | | | | | |
| $\overline{X}_1$ | | | | | | |

*Indicates significance for data in table 2.

### 4.5 Duncan's New Multiple Range Test - $\alpha_p = .05$

To find the critical value for the $\overline{X}_6 - \overline{X}_1$ cell, find $q^*(.05, 6-1+1, 24) = 3.28$ in a table of critical values for Duncan's new multiple range test and multiply by $s/\sqrt{n}$, so that $3.28(22.14) = 72.6$.

To find the critical value for the $\overline{X}_4 - \overline{X}_1$ cell, find $q^*(105, 4-1+1, 24) = 3.16$ in the same table and multiply by $s/\sqrt{n}$ so that $3.16(22.14) = 70.0$.

The test was concluded when $\overline{X}_6 - \overline{X}_5$, $\overline{X}_5 - \overline{X}_2$, and $\overline{X}_4 - \overline{X}_1$, were all found nonsignificant. The results are again summarized as in 4.1.

$$\begin{array}{cccccc} \overline{X}_1 & \overline{X}_2 & \overline{X}_3 & \overline{X}_4 & \overline{X}_5 & \overline{X}_6 \\ \cline{1-4} \cline{5-6} \\[-2ex] \cline{2-5} \end{array}$$

Table 4 --Critical Values for Duncan's Method with $\alpha = .05$, $K = 6$, $v = 24$

| | $\overline{X}_1$ | $\overline{X}_2$ | $\overline{X}_3$ | $\overline{X}_4$ | $\overline{X}_5$ | $\overline{X}_6$ |
|---|---|---|---|---|---|---|
| $\overline{X}_6$ | 72.6* | 71.5* | 70.0* | 68.0* | 64.6 | |
| $\overline{X}_5$ | 71.5* | 70.0 | | | | |
| $\overline{X}_4$ | 70.0 | | | | | |
| $\overline{X}_3$ | | | | | | |
| $\overline{X}_2$ | | | | | | |
| $\overline{X}_1$ | | | | | | |

*Indicates significant cells for data in table 2.

## 5. TESTING MULTIPLE CONTRASTS AMONG TREATMENT MEANS

Many times an experimenter is interested in "multiple" contrasts, which are differences between groups of means. The data may suggest relationships among treatment means that previously were not apparent, or the experimenter may wish to test some multiple contrasts for which independent sums of squares cannot easily be partitioned in the analysis of variance table.

A linear contrast among $k$ true treatment means is defined as

$$\psi = \sum_{m=1}^{k} C_m\mu_m = C_1\mu_1+C_2\mu_2+\ldots+C_k\mu_k,$$

where the $C_m$'s are any constants subject to the condition:

$$\sum_{m=1}^{k} C_m = C_1 + C_2 +\ldots+ C_k = 0.$$

One possible multiple linear contrast among six means is

$$\psi = (\mu_1 + \mu_5)/2 - (\mu_2 + \mu_4 + \mu_6)/3,$$

where the coefficients, $C_1 = 1/2$, $C_2 = -1/3$, $C_3 = 0$, $C_4 = -1/3$, $C_5 = 1/2$, $C_6 = -1/3$, sum to zero.

Differences between pairs of treatment means are actually simplified forms of multiple contrasts where each group of means contains only one member. In notation,

$$\mu_i - \mu_j = \sum_{m=1}^{k} C_m\mu_m, \text{ where } C_m = \begin{matrix} 1, \text{ if } m = i \\ -1, \text{ if } m = j \\ 0, \text{ otherwise} \end{matrix} \quad \text{and} \quad \sum_{m=1}^{k} C_m = 1+(-1) = 0$$

Methods for simultaneously testing such "simple" linear contrasts are described in sections 2, 3, and 4.

There are two general methods of placing simultaneous confidence intervals around multiple contrasts among treatment means (Scheffé 1959). Both provide an experimentwise error rate.

### 5.1 S-Method (Scheffé) - $\alpha_w$ Error Rate

Under the usual assumptions ($\overline{X}_1$, ..., $\overline{X}_k$ are independently, identically, normally distributed random variables), the probability is 1-α that simultaneously for all possible independent linear contrasts,

$$\psi = \sum_{m=1}^{k} C_m\mu_m, \quad \sum_{m=1}^{k} C_m\overline{X}_m - S(\sum_{m=1}^{k} C_m^2)^{1/2}s/\sqrt{n} \leq \psi \leq \sum_{m=1}^{k} C_m\overline{X}_m + S(\sum_{m=1}^{k} C_m^2)^{1/2}s/\sqrt{n},$$

where

$$S = \{(k-1)F\alpha,\ k-1,\ v\}^{1/2},$$

$s$ is the estimate of σ based on $v$ degrees of freedom, and each experimental mean is based upon $n$ observations.

### 5.2 T-Method (Tukey) - $\alpha_w$ Error Rate

Under the same assumptions as in 5.1, the probability is 1-α that simultaneously for all contrasts,

$$\psi = \sum_{m=1}^{k} C_m\mu_m, \quad \sum_{m=1}^{k} C_m\overline{X}_m - q(\alpha, k, v)s/\sqrt{n} \cdot 1/2 \sum_{m=1}^{k} |C_m| \leq \psi \leq \sum_{m=1}^{k} C_m\overline{X}_m + q(\alpha, k, v)s/\sqrt{n} \cdot 1/2 \sum_{m=1}^{k} |C_m|,$$

where $q(\alpha, k, v)$ is the upper 100α percent point of the Studentized range based on $k$ means and $v$ degrees of freedom, $s$ and $n$ are as in 5.1, and $|C_m|$ is the absolute value of $C_m$.

Notice that, if the contrast is a simple linear contrast,

$$1/2 \sum_{m=1}^{k} |C_m| = 1/2 \ \{|1| + |-1|\} = 1$$

so that the interval above is precisely that given by Tukey's method of section 3.3 .

## 5.3 Discussion

These confidence intervals may be used to provide a test of hypothesis. Reject the null hypothesis,

$$H_o:\psi = \sum_{m=1}^{k} C_m\mu_m = 0,$$

if the confidence interval does not contain 0, and accept the null hypothesis if the confidence interval does contain 0.

As pointed out in section 3.3, the *T*-method is more powerful for testing simple linear contrasts. It has been shown that the *S*-method is more powerful for testing multiple linear contrasts (Scheffé 1959). Thus a good suggestion is to use the *T*-method if only simple linear contrasts are to be tested or if a mixture of simple and multiple linear contrasts are to be tested. Use the *S*-method if only independent multiple linear contrasts are to be tested.

The *S*-method is valid for testing either simple or multiple contrasts among adjusted means from an analysis of covariance (Halperin and Greenhouse 1958).

Sometimes the assumption of independence among treatment means is not reasonable. But if the experimenter believes that all the covariances between means are equal, an extension of the *T*-method may be used to test linear contrasts among the means (Scheffé 1959).

## 5.4 Example

Consider the data presented in section 4. Suppose treatments 3, 5, 6 were administered in the presence of an impurity and treatments 1, 2, 4 were free of the impurity. The experimenter may feel that it is important to test the multiple contrast

$$\psi = \mu_1 + \mu_2 - \mu_3 + \mu_4 - \mu_5 - \mu_6.$$

*S*-method, $\alpha_w = .05$

$$\sum_{m=1}^{k} C_m\overline{X}_m = 1470 + 1498 - 1505 + 1528 - 1564 - 1600$$

$$= -173$$

$$\sum_{m=1}^{k} C_m^2 = 1^2 + 1^2 + (-1)^2 + 1^2 + (-1)^2 + (-1)^2 = 6$$

$$F(.05,\ 5,\ 24) = 2.62$$

$$s/\sqrt{n} = 22.14$$

$$S = \{5(2.62)\}^{1/2} = 3.62$$

Therefore,

$$Pr\ \{-173 - (3.62)(6)^{1/2}(22.14) \leq \psi \leq -173 + (3.62)(6)^{1/2}(22.14)\} = .95$$

or

$$Pr\ \{-369.4 \leq \psi \leq 23.4\} = .95$$

*T*-Method, $\alpha_w = .05$

$$\sum_{m=1}^{k} C_m\overline{X}_m = -173$$

$$\sum_{m=1}^{k} |C_m| = 1 + 1 + |-1| + 1 + |-1| + |-1| = 6$$

$$q(.05,\ 6,\ 24) = 4.37$$

$$s/\sqrt{n} = 22.14$$

Therefore,

$$Pr\ \{-173 - (4.37)(22.14)(6/2) \leq \psi \leq -173 + (4.37)(22.14)(6/2)\} = .95$$

or

$$Pr\ \{-463.3 \leq \psi \leq 117.3\} = .95$$

The results of both methods suggest that the average of means of treatments containing the impurity were not significantly different from the average of means of treatments not containing the impurity.

This example shows that confidence intervals about a multiple linear contrast can be much wider with the *T*-method than with the *S*-method.

## LITERATURE CITED

Bose, R. C. and Roy, S. N.
1953. Simultaneous confidence interval estimation. Ann. Math. Statis. 24: 513.

Duncan, D. B.
1955. Multiple range and multiple F tests. Biometrics 11: 1.

Dunn, Olive J.
1961. Multiple comparisons among means. Amer. Statis. Assoc. Jour. 56: 52.

Dunnett, C. W.
1955. A multiple comparison procedure for comparing several treatments with a control. Amer. Statis. Assoc. Jour. 50: 1096.

Dwass, M.
1959. Multiple confidence procedures. Ann. Inst. Statis. Math. 10: 277.

Federer, W. T.
1955. Experimental design: Theory and application. 544 pp., New York: McMillan Co.

Halperin, M. and Greenhouse, S. W.
1958. Note on multiple comparison for adjusted means in the analysis of covariance. Biometrika 45: 256.

Harter, H. L.
1957. Error rates and sample sizes for range tests in multiple comparisons. Biometrics 13: 511; Correction, 1961, Biometrics 17: 321.

__________
1960a. Tables of range and studentized range. Ann. Math. Statis. 31: 1122.

__________
1960b. Critical values for Duncan's new multiple range test. Biometrics 16: 671.

Hartley, H. O.
1955. Some recent developments in analysis of variance. Commun. on Pure and Applied Math. 8: 47.

Kramer, C. Y.
1956. Extension of multiple range tests to group means with unequal number of replications. Biometrics 12: 307.

Roy, S. W.
1954. Some further results in simultaneous confidence interval estimation. Ann. Math. Statis. 25: 752.

Sarhan, A. E. and Greenberg, B. G.
1962. Contributions to order statistics. 482 pp., New York: Wiley and Sons, Inc.

Scheffé, H.
1959. The analysis of variance. 477 pp., New York: Wiley and Sons, Inc.

Agriculture --- CSU, Ft. Collins